The following notice applies to any unclassified (including originally classified and now declassified) technical reports released to "qualified U.S. contractors" under the provisions of DoD Directive 5230.25, Withholding of Unclassified Technical Data From Public Disclosure.

NOTICE TO ACCOMPANY THE DISSEMINATION OF EXPORT-CONTROLLED TECHNICAL DATA

1. Export of information contained herein, which includes, in some circumstances, release to foreign nationals within the United States, without first obtaining approval or license from the Department of State for items controlled by the International Traffic in Arms Regulations (ITAR), or the Department of Commerce for items controlled by the Export Administration Regulations (EAR), may constitute a violation of law.

2. Under 22 U.S.C. 2778 the penalty for unlawful export of items or information controlled under the ITAR is up to ten years imprisonment, or a fine of $1,000,000, or both. Under 50 U.S.C., Appendix 2410, the penalty for unlawful export of items or information controlled under the EAR is a fine of up to $1,000,000, or five times the value of the exports, whichever is greater; or for an individual, imprisonment of up to 10 years, or a fine of up to $250,000, or both.

3. In accordance with your certification that establishes you as a "qualified U.S. Contractor", unauthorized dissemination of this information is prohibited and may result in disqualification as a qualified U.S. contractor, and may be considered in determining your eligibility for future contracts with the Department of Defense.

4. The U.S. Government assumes no liability for direct patent infringement, or contributory patent infringement or misuse of technical data.

5. The U.S. Government does not warrant the adequacy, accuracy, currency, or completeness of the technical data.

6. The U.S. Government assumes no liability for loss, damage, or injury resulting from manufacture or use for any purpose of any product, article, system, or material involving reliance upon any or all technical data furnished in response to the request for technical data.

7. If the technical data furnished by the Government will be used for commercial manufacturing or other profit potential, a license for such use may be necessary. Any payments made in support of the request for data do not include or involve any license rights.

8. A copy of this notice shall be provided with any partial or complete reproduction of these data that are provided to qualified U.S. contractors.

DESTRUCTION NOTICE

For classified documents, follow the procedure in DoD 5220.22-M, National Industrial Security Program, Operating Manual, Chapter 5, Section 7, or DoD 5200.1-R, Information Security Program Regulation, Chapter 6, Section 7. For unclassified, limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI[1]

Joseph Polifroni, Stephanie Seneff, and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

As part of our development of a spoken language system in the ATIS domain, we have begun a small-scale effort in collecting spontaneous speech data. Our procedure differs from the one used at Texas Instruments (TI) in many respects, the most important being the reliance on an existing system, rather than a wizard, to participate in data collection. Over the past few months, we have collected over 3,600 spontaneously generated sentences from 100 subjects. This paper documents our data collection process, and makes some comparative analyses of our data with those collected at TI. The advantages as well as disadvantages of this method of data collection will be discussed.

INTRODUCTION

ATIS, or Air Travel Information Service, is the designated common task of the DARPA Spoken Language Systems (SLS) Program [8]. As part of our development of a spoken language system in this domain, we have recently begun a small-scale effort in collecting spontaneous speech data. This effort is motivated partly by our desire to contribute to the data collection efforts already underway elsewhere [4,2,1], so that more data can be available to the community sooner for system development, training, and evaluation. In addition, we were interested in exploring various alternatives of the data collection procedure itself. It is our belief that we as a community do not fully understand how goal-directed spontaneous speech should best be collected. This is not surprising, since we have little experience in this area. Nevertheless, data collection is an important area of research for the SLS Program, since the type of data that we collect will directly affect the capabilities of systems that we develop, and the evaluations that we can perform. Therefore, we thought it would be appropriate to experiment with different aspects of this process. There is evidence that even very small changes in the procedure, such as the instructions to the subject, can drastically alter the nature of the data collected [1].

The paper is organized as follows. We will first discuss some methodological considerations that led to the particular collection procedure that we adopted. We will then briefly describe the procedure itself. This will be followed by some comparative analyses of a subset of the data that we have collected with those collected at Texas Instruments (TI). Implications of our findings will be discussed.

DATA COLLECTION

As is the case with other efforts [4,2,1], our data are collected under simulation. Nevertheless, we wanted the simulation to reflect as much as possible the system that we are developing. In this section, we will briefly describe some design issues and document the actual collection process. Further details can be found elsewhere [7].

Methodological Considerations

While many years may pass before we are able to build systems with capabilities approaching those of humans, we believe strongly that it should soon be possible to develop functioning systems with limited capabilities. The successful development of such systems will partly depend on our ability to train *subjects* to stay within the restricted domain of the system. Therefore, we should try to collect data intentionally restricting the user in ways that closely match system capability. In this section we will describe some aspects of our data collection paradigm that support this viewpoint.

Wizard *vs.* System By far the most important difference between the data collection procedures at TI and MIT is the way system simulation is conducted during data collection. TI made use of a "wizard" paradigm, in which a highly skilled experimenter interprets what was spoken, converts it into a form that enables database access, and produces an answer for the subject [4,2]. Based on our previous positive experience with collecting spontaneous speech for a different domain [10], we decided to explore an alternative paradigm from the one used at TI, in which we make use of the system under development to do most of the work. That is, prior to the beginning of data collection, the natural language component is developed to the point where it has reasonable coverage of

---

[1] This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **1991**	2. REPORT TYPE	3. DATES COVERED **00-00-1991 to 00-00-1991**

4. TITLE AND SUBTITLE **Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Computer Science,Spoken Language Systems Group,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**6**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

```
Subject: Show flights from Philadelphia to Denver serving lunch or dinner on February second and also show their fares.
ATIS Response:
------------------------------------------------------------------------------------------------------------------------
(RAW DISPLAY)                                                   | (PROCESSED DISPLAY)
                                                                | These are the flights from Philadelphia to Denver serving
                                                                | lunch and dinner on Friday February 2.
AIRLINE FLIGHT   FROM     TO    DEPARTURE ARRIVAL STOPS MEAL    | AIRLINE NUMBER FROM TO  DEPARTURE   ARRIVAL STOPS MEALS
 CODE   NUMBER  AIRPORT AIRPORT   TIME      TIME        CODE    |                                                SERVED
 DL     1083     PHL     DEN     1230      1615      1   LS     | DELTA    1083   PHL  DEN 12:30 P.M. 4:15 P.M.  1  LUNCH SNACK
 UA     355      PHL     DEN     1753      1955      0   D      | UNITED  355     PHL  DEN  5:53 P.M. 7:55 P.M.  0  DINNER
 CO     1631     PHL     DEN     1800      2037      0   D      | CONT     1631   PHL  DEN  6:00 P.M. 8:37 P.M.  0  DINNER
------------------------------------------------------------------------------------------------------------------------
(RAW DISPLAY)                                                   | (PROCESSED DISPLAY)
                                                                | These are the fares for the flights from Philadelphia to Denver
                                                                | serving lunch and dinner on Friday February 2.
AIRLINE  FLIGHT    RESTRICT  ONE WAY  RND TRIP   FARE           | AIRLINE NUMBER RESTRICTION  ONE WAY ROUND TRIP    FARE CLASS
CODE     NUMBER    CODE      COST     COST       CLASS          |
 CO      1631                 470       940       Y             | CONT     1631       NONE    $470.00   $940.00    Y: COACH CLASS
 CO      1631                 706      1412       Y             | CONT     1631       NONE    $706.00  $1412.00    F: FIRST CLASS
 DL      1083                 420       840       Y             | DELTA    1083       NONE    $420.00   $840.00    Y: COACH CLASS
 DL      1083                 630      1260       F             | DELTA    1083       NONE    $630.00  $1260.00    F: FIRST CLASS
 UA      355                  470       940       Y             | UNITED   355        NONE    $470.00   $940.00    Y: COACH CLASS
 UA      355                  706      1412       F             | UNITED   355        NONE    $706.00  $1412.00    F: FIRST CLASS
```

Figure 1: Comparison of the displays as returned from the OAG database (left panels) and those presented to the subject (right panels) for a query.

the possible queries. In addition, the system must be able to automatically translate the text into a query to the database, and return the information to the subject. Once such a system is available, data collection is accomplished by having the experimenter, a fast and accurate typist, type verbatim to the system what was spoken, after removing spontaneous speech disfluencies. The actual interpretation and response generation is accomplished by the system without further human intervention. If the sentence cannot be understood by the system, an error message is produced to help the subject make appropriate modifications.

Another feature of our paradigm is that the underlying system can be improved incrementally using the data collected thus far. The resulting expansion in system capabilities permit us to accommodate more complex sentences as well as those that previously failed.

Displays One of the considerations that led to the selection of ATIS as the common task is the realization that, since most people have planned air travel at one time or another, there will be no shortage of subjects familiar with the task. Since the average traveller is not likely to be knowledgeable of the format and display of the Official Airline Guide (OAG), we have translated many of the cryptic symbols and abbreviations that OAG uses into easily recognizable words. We believe that this change has the positive effect of helping the subject focus on the travel planning aspect, and not be confused by the cryptic displays that are intended for more experienced users. In fact, we try to keep the displayed information at a minimum in order to encourage verbal problem solving. In general, we only display the airline, flight number, origination and destination cities, departure and arrival time, and the number of stops. Additional columns of information are included only when specifically requested.

Figure 1 illustrates the difference between the raw displays returned from the OAG database and the ones that we present to the subjects by applying some post-processing to the raw display. The query is "Show flights from Philadelphia to Denver serving lunch or dinner on February second and also show their fares." Note that the airlines, meal codes, and fare codes in the processed displays (shown in the right-hand panels) have all been translated into words as much as possible while keeping the displays manageable on a screen.[2] The military-time displays for departure and arrival times have also been converted to more familiar forms to facilitate interpretation. Furthermore, under the TI data collection scheme the answer is assembled as one large table. However, we break it up into two answers, one for the flights and one for the fares.

System Feedback Our system provides explicit feedback to the subject in the form of text and synthetic speech, paraphrasing its understanding of the sentence. This feature is illustrated in Figure 1 in the right-hand panels, immediately above the display tables. By providing confirmation to the subject of what was understood, the system greatly reduces the confusion and frustration that may arise later on in the dialogue caused by an earlier error in the system's responses. In addition, the generation of a verbal response implicitly encourages the notion of human/machine interactive dialogue.

Interactive Dialogue We believe that, for the ATIS system to be truly useful, the user must be able to carry out an

---

[2] Some of the headings in the raw display have been reformatted so that the table will stay within the width of this page.

interactive dialogue with the system, in the same way that a traveller would with a travel agent. Our data collection procedure therefore encourages natural dialogue by incorporating some primitive discourse capabilities, allowing subjects to make indirect as well as direct anaphoric references, and fragmentary responses where appropriate. In some sessions, we even use a version of the system that plays an active role in guiding the subject through flight reservations. Details of the interactive dialogue capabilities of our system are described in a companion paper [9].

Data Collection Process

The data are collected in an office environment where the ambient noise is approximately 60 dB SPL, measured on the C scale. The subject sits in front of a display monitor, with a window slaved to a Lisp process running on the experimenter's Sun workstation located in a nearby room. The experimenter's typing is hidden from the subject to avoid unnecessary distractions. A push-to-talk mechanism is used to collect the data. The subject is instructed to hold down a mouse button while talking, and release the button when done. The resulting speech is captured both by a Sennheiser HMD-224 noise-cancelling microphone and a Crown PZM desk-top microphone, digitized simultaneously.

Prior to the session, the subject is given a one-page description of the task, a one-page summary of the system's knowledge base, and three sets of scenarios [7]. The first set contains simple tasks such as finding the earliest flight from one city to another that serves a meal, and is intended as a "warm-up" exercise. The second set involves more complex tasks, and includes the official ATIS scenarios. Finally, the subject is asked to make up a scenario and attempt to solve it. The subjects are instructed to choose a pre-determined number of scenarios from each category. In addition, they are asked to clearly delineate each scenario with the commands "begin scenario x" and "end scenario x," where x is a number, so that we can keep track of discourse utilization. A typical session lasts about 40 minutes, including initial task familiarization and final questionnaire.

For several initial data collection sessions, the first two authors took turns serving as the experimenter. Once we began daily sessions, however, it was possible to hire a part-time helper to serve as the scheduler, experimenter, and transcriber. The experimenter can hear everything that the subject says and can communicate with the subject via a two-way microphone/speaker hook-up. However, the experimenter rarely communicates with the subject once the experiment is under way. The digitized speech is played back to the experimenter, allowing him/her to confirm that the recording was successful. The voice input during the session, minus disfluencies, is typed verbatim to ATIS by the experimenter, and saved automatically in a computer log. The system response is generated automatically from this text, and is also recorded into the log. The system's response typically takes less than 10 seconds after the text has been entered. At a later time, the experimenter listens again to each digitized sentence and inserts false starts and non-speech events into the orthography to form a detailed orthographic transcription, following the conventions described in [6].

There are basically three ways that the system can fail, each of which provides a distinct error message. If the sentence contains unknown words, then the system identifies to the subject the words that it doesn't know. If it knows all the words, but the sentence fails to parse, then it identifies the point in the sentence where the parse failed. Finally, it may parse the sentence but fail to produce a response due to, for instance, an incorrect database query. In that case, it simply says, "I ran into an error trying to evaluate this sentence." Our long-term goal is to make error messages sufficiently informative that the subject knows how to best rephrase the query. By examining how subjects react to the various kinds of error messages, we hope to improve the overall usability of the system. Figure 2 illustrates the data collection process with a simple dialogue between a subject and ATIS.

```
SCENARIO: "Find the earliest (or latest) flight from one
city to another that serves a meal of your choice."

Subject: BEGIN SCENARIO TWO
Subject: SHOW ME THE FLIGHTS FROM BOSTON TO ATLANTA
Atis: These are the flights from Boston to Atlanta.
 AIRLINE NUMBER FROM TO  DEPARTURE   ARRIVAL   STOPS
 DELTA    225   BOS  ATL  6:30 A.M   9:13 A.M.   0
 EASTERN  147   BOS  ATL  7:00 A.M.  9:46 A.M.   0
 DELTA    219   BOS  ATL  8:15 A.M. 11:00 A.M.   0
 EASTERN  643   BOS  ATL  8:32 A.M. 11:26 A.M.   0
 DELTA    547   BOS  ATL  9:55 A.M. 12:35 P.M.   0
 EASTERN  145   BOS  ATL 10:35 A.M.  1:27 P.M.   0
 DELTA    708   BOS  ATL 11:56 A.M.  2:35 P.M.   0
 EASTERN  143   BOS  ATL 12:40 P.M.  3:25 P.M.   0
 EASTERN  645   BOS  ATL  2:21 P.M.  5:09 P.M.   0
 DELTA    975   BOS  ATL  3:20 P.M.  6:04 P.M.   0
 EASTERN  141   BOS  ATL  4:05 P.M.  7:00 P.M.   0
 EASTERN  149   BOS  ATL  5:58 P.M.  8:49 P.M.   0
 DELTA    323   BOS  ATL  6:50 P.M.  9:28 P.M.   0
 DELTA    54    BOS  ATL  8:24 P.M. 10:58 P.M.   0

Subject:  PLEASE SHOW ME THE THREE EARLIEST FLIGHTS
Atis: I'm sorry I got stuck just before the words
      "earliest flights".

Subject: WHICH ONE IS THE EARLIEST FLIGHT
Atis: The earliest flight from Boston to Atlanta?
 AIRLINE NUMBER FROM TO  DEPARTURE ARRIVAL  STOPS
 DELTA   225    BOS  ATL 6:30 A.M  9:13 A.M. 0

Subject: DOES THIS FLIGHT SERVE BREAKFAST
Atis: Yes, the earliest flight from Boston to Atlanta does
      serve breakfast.
 AIRLINE NUMBER FROM TO  DEPARTURE ARRIVAL  STOPS  MEALS
 DELTA   225    BOS  ATL 6:30 A.M. 9:13 A.M.  0   BREAKFAST

Subject: END SCENARIO TWO
```

Figure 2: An example log for one scenario created by a subject

Subjects are recruited from the general vicinity of MIT. No restrictions in age or sex are imposed, nor do we insist that they be native speakers. For their efforts, each subject is given a gift certificate at a popular Chinese restaurant or an ice-cream parlor. Presently, we are collecting data from two to three subjects per day.

COMPARATIVE ANALYSES

To facilitate system development, training, and testing, we arbitrarily partitioned part of the collected data into training, development-test, and test sets, as summarized in Table 1. All the comparative analyses reported in this section are based on our designated training set and the TI training set, the latter defined as the total amount of training data released by TI prior to June 1990.

Data Set	# Speakers	# Sentences
Training	41	1582
Development-Test	10	371
Test	10	324

Table 1: Designation of various data sets collected at MIT.

General Characteristics Table 2 compares some general statistics of the data in the TI and MIT training set. On the average, the wizard paradigm used at TI can collect 25 sentences over approximately 40 minutes, for a yield of 39 sentences per hour [2]. In contrast, we were able to collect an average of about 39 sentences in approximately 45 minutes, for a yield of 53 sentences per hour. Our higher yield is presumably due to the fact that the system can respond much faster than a wizard; the process of translating the sentences into an NLParse command [2] by hand can sometimes be quite time-consuming. Note that the yields in both cases do not include the generation of the ancillary files, which is an essential task performed after data collection.

Variables	TI Data	MIT Data
# Speakers	31	41
# Sentences	774	1582
Ave. # Sentences/Speaker	25.0	38.6
Ave. # Words/Sentence	10.65	9.14
% of Table Clarification Sentences	25	1
Ave. # Words/Second	1.18	2.04

Table 2: General statistics of the training data collected at TI and MIT.

The average number of words per sentence for the MIT data is 15% fewer than that for the TI data. The shorter sentences in the MIT data can be due to several reasons. The system's inability to deal with longer sentences and the feedback that it provides may coerce the subject into making shorter sentences. The limited display may discourage the construction of lengthy and sometimes contorted sentences that attempt to solve the scenarios expeditiously. The interactive nature of problem solving may encourage the user to take a "divide-and-conquer" attitude and ask simpler questions. Closer examination of the data reveals that the standard deviation on sentence length is very different between the two data sets ($\sigma_{TI} = 5.53$ and $\sigma_{MIT} = 3.68$). We suspect that this is primarily due to the preponderance of short symbol clarification sentences such as "What does EA mean?" in the TI data, along with occasional very long sentences. Table 2 shows that 25% of the TI sentences deal with table clarification compared to only 1% of our sentences. In fact, 8 of our 16 table clarification sentences concern airline code abbreviations. They were collected from earlier sessions when the display was still somewhat cryptic. Once we made some extremely simple changes in the display, such sentences no longer appeared.

The speaking rate of the MIT sentences was more than 70% higher that that of the TI sentences. We believe that the speaking rate of the TI sentences (70 words/minute) is unnaturally low. This may be due to the insertion of many pauses, or the fact that the subjects simply spoke tentatively, due to their unfamiliarity with the task. Acoustic analysis is clearly needed before we can know for certain.

System Growth Rate Figure 3 compares the size of the lexicon, i.e., the number of unique words, as a function of the number of training sentences collected at TI and MIT. The Figure shows that the vocabulary size grows at a much slower rate (about 20 words per 100 training sentences) for the MIT ATIS data than the TI data (about 50 words per 100 training sentences). Also included on the Figure for reference is a plot of the growth rate for our VOYAGER corpus, which was collected using the same paradigm as we have used for ATIS. A previous comparison of the TI data and the MIT VOYAGER data [5] led to the conclusion that the VOYAGER *domain* was intrinsically more restricted. Since the MIT ATIS data are more similar to the VOYAGER data, it may be the case that a more critical factor was the data collection paradigm. Thus, one may argue that our data collection procedure is better able to encourage the subjects to stay within the domain. A slow growth rate may also be an indication that the training data is more representative of the unseen test data.

As further evidence that our training data is representative of the data that the system is likely to see, Table 3 compares the system's performance on the MIT training and development-test sets. The similarities in performance between the two data sets is striking, suggesting that the system is able to generalize well from training data. Since the system can deal with over 70% of the sentences, we feel that the subject is not likely to be overly frustrated by the system's inability to deal with the remaining sentences. This also reflects the apparent ability of subjects to adjust their speech so as to stay generally within the domain of the system.

Disfluencies Table 4 compares the occurrence of spontaneous speech disfluencies in the two data sets. We define

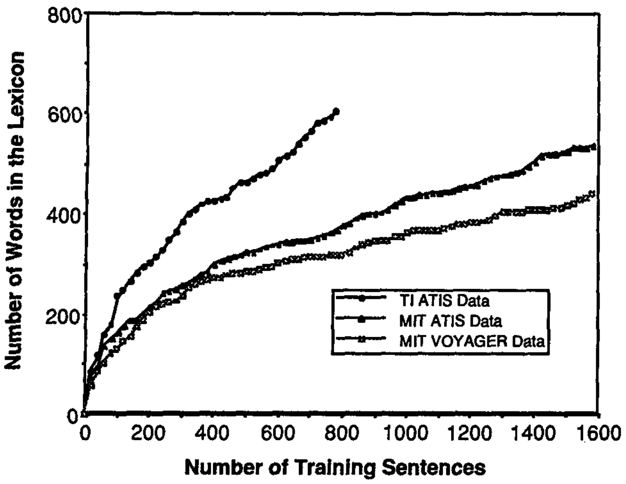

Figure 3: The size of the lexicon as a function of the number of training sentences for the TI and MIT ATIS training sets, as well as the MIT VOYAGER training set.

% of Sentences	Training Set	Development-Test Set
with New Words	9.3	8.6
with NL failure	16.9	19.7
Parsed	73.8	71.7

Table 3: System performance on the training and development-test sets. Evaluation on the training set was conducted after the system had been trained on these sentences, whereas the development-test set represents unseen data.

% of Sentences	TI Data	MIT Data
with filled pauses	8.1	1.3
with lexical false starts	6.0	2.8
with linguistic false starts	5.9	1.0

Table 4: Analyses of disfluencies in the training data collected at TI and MIT.

lexical false starts as the appearance of a partial word and linguistic false starts as the appearance of one or more extraneous whole words. Again, our analyses show quite a difference between the two data sets along all dimensions. A total of 73 filled pauses appear in 63 (or 8.1%) of the TI sentences, whereas only 25 appear in 21 (or 1.3%) of the MIT sentences. Similarly, it is twice as likely to find a sentence with a lexical false start in the TI data as is in the MIT data, and almost six times more likely for a linguistic false start.

DISCUSSION

By far the most important feature of our data collection process is that the system under development is used in place of a wizard. We believe that this "system-assisted" paradigm offers several advantages. Since the system used to collect the data is under continual development, we can periodically replace it with an improved version, where the improvement will be guided by the data already collected. If, for example, we observe that subjects frequently use a certain linguistic construct, then we can modify the system to provide that capability. Alternatively, we may decide that the capability is outside the domain of expertise of the system (for example, booking flights). We can then try to keep the subject "in bounds" by experimenting with different subject instructions. Furthermore, this type of data collection provides relatively realistic sample data of human-machine interaction. This distinguishes it from wizard data, where the human wizard will answer any query that the back-end can answer. We believe that by providing the subject with a more realistic situation, where there are a significant number of queries that the system cannot handle, we can gather better data about the possible training effects of the system on the subject, the subject's ability to adapt to the system, and the system capabilities to provide useful diagnostics on its limitations. By combining data collection and system development into closely coupled cycles, we can potentially ensure that the type of data that we collect is appropriate for the system that we want to develop, thus increasing the efficiency of system development.

Our data collection procedure is also cost effective. Since the experimenter plays a very passive role during data collection, we eliminate the need for a highly skilled, and presumably expensive, person to interpret what was said and coerce the back-end to generate the necessary responses. This has the dual effect of freeing the researchers to concentrate on system development, and reducing the cost of data collection. We estimate that the unburdened cost for data collection, including the subject, the experimenter, and the generation of the correct transcriptions, is about $0.85 per sentence. Subsequent categorization of the sentences, and the generation of the reference answers will add another $2.30 to each sentence, although this may not be needed for all the data collected.

The disadvantage of this method of data collection is that the baseline system is constantly evolving. This has several effects. First, data collected in an earlier session may not be comparable to data collected in a later session. This is not important if one merely wishes to collect as much spontaneous speech within a given domain as possible. However, it can create a consistency problem if the data are used for training at a later stage. For example, in an earlier session, the system may provide an error message stating that it does not understand a particular word, whereas in a later session it may be able to handle the identical query with no problem. The system response can also change from a diagnostic message to a request for further information or clarification. The result is that a dialogue collected at one stage in system development may not be coherent at a later stage in system development. This is a problem for development of dialogue handling. However, it may be possible to handle this by evaluating the system after "resetting the context" based on the actual system response, as suggested in [3]. Despite these difficulties, we feel that the ease of data collection with this methodology by far outweighs any disadvantages.

Comparative analyses of the data collected at TI and MIT reveal significant differences in many dimensions. Given the many ways in which the two procedures differ, it is not always easy to attribute the discrepancies to one single factor. One of the most striking differences between the TI and MIT data is the fraction of sentences dealing with table clarification. One may argue that these sentences are unnecessary by-products of the cryptic display format, and they contribute very little to the problem of providing graceful human/machine interface for travel planning. By a very simple change in the display format, we were able to reduce this type of question by 25 fold! Similarly, a change in data collection procedure can reduce the number of spontaneous speech phenomena several fold. It is therefore conceivable that we can minimize the occurrence of spontaneous speech phenomena in *real* systems. These effects again underscore the importance of collecting the type of data that is as closely matched as feasible to the capabilities of the system that we are developing.

While we have used our own natural language system, TINA, for data collection, it is important to note that the choice was primarily motivated by convenience and availability. Clearly, any functioning natural language system could be freely substituted. In fact, a richer pool of data would probably arise if data were collected independently at several sites, each of which used *their own* system for the back-end responses.

At this writing, we have collected 3,690 sentences from 102 subjects. Orthographic transcriptions for all of the sentences are available, as are the categorizations and reference answers for the development-test and test sets.

ACKNOWLEDGEMENT

We would like to acknowledge the help of Claudia Sears, who served as experimenter and transcriber for much of the data, Victoria Palay, who helped in the coordination of the data collection, and Bridget Bly of SRI, whom we hired as a consultant in order to generate some of the ancillary files for the development-test and test sets.

REFERENCES

[1] Bly, B., P. Price, S. Tepper, E. Jackson, and V. Abrash, "Designing the Human Machine Interface in the ATIS Domain," *Proc. Third Darpa Speech and Language Workshop*, pp. 136-140, Hidden Valley, PA, June 1990.

[2] Hemphill, C. T., J. J. Godfrey, and G. R. Doddington, "The ATIS Spoken Language System Pilot Corpus," *Proc. Third Darpa Speech and Language Workshop*, pp. 96-101, Hidden Valley, PA , June 1990.

[3] Hirschman, L., D. A. Dahl, D. P. McKay, L. M. Norton, L., and M. C. Linebarger, "Beyond Class A: A Proposal for Automatic Evaluation of Discourse," *Proc. Third Darpa Speech and Language Workshop*, pp. 109-113,Hidden Valley, PA , June 1990.

[4] Kowtko, J. C. and P. J. Price, "Data Collection and Analysis in the Air Travel Planning Domain," *Proc. Second Darpa Speech and Language Workshop*, pp. 119-125, Harwichport, MA, October 1989.

[5] Norton, L. M., D. A. Dahl, D. P. McKay, L. Hirschman, M. C. Linebarger, D. Magerman, and C. N. Ball, "Management and Evaluation of Interactive Dialog in the Air Travel Domain," *Proc. Third Darpa Speech and Language Workshop*, pp. 141-146, Hidden Valley, PA , June 1990.

[6] Polifroni, J. and M. Soclof, "Conventions for Transcribing Spontaneous Speech Events in the VOYAGER Corpus," DARPA SLS Note 6, Spoken Language Systems Group, MIT Laboratory for Computer Science, Cambridge, MA, February, 1990.

[7] Polifroni, J., S. Seneff, V. W. Zue, and L. Hirschman, , "ATIS Data Collection at MIT," DARPA SLS Note 8, Spoken Language Systems Group, MIT Laboratory for Computer Science, Cambridge, MA, November, 1990.

[8] Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third Darpa Speech and Language Workshop*, pp. 91-95, Hidden Valley, PA , June 1990.

[9] Seneff, S., L. Hirschman, and V. W. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," *These Proceedings*.

[10] Zue, V., N. Daly, J. Glass, H. Leung, M. Phillips, J. Polifroni, S. Seneff, and M. Soclof, "The Collection and Preliminary Analysis of a Spontaneous Speech Database," *Proc. Second Darpa Speech and Language Workshop*, pp. 126-134, Harwichport, MA , October 1989.